International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 5, Sept-Oct-2021, ISSN (Online): 2395-566X

# Comparative Analysis of Balancing Methods for Classifying Imbalanced Data

**Himani Tiwari, Dr. Sheetal Rathi**
Dept. Of Computer Engineering
Thakur College Of Engineering And Technology
Mumbai, India

**Abstract- The classification of data with unbalanced class distribution encounters the significant shortcomings of the performance that most standard classification learning algorithms can achieve. These algorithms assume that the class distribution is relatively balanced and the cost of pre-classification is the same. This article reviews the classification of unbalanced data: areas of application; the nature of the problem; learning difficulties with standard classification learning algorithms; learning objectives and evaluation measures; reported research solutions and class imbalance problems when there are multiple classes.**

**Keywords- Classification; class imbalance problem.**

## I. INTRODUCTION

Classification is an important task for pattern recognition. A number of classification learning algorithms, such as decision trees, neural networks in backpropagation, Bayesian networks, nearest neighbors, support vector machines, and recently reported association classifications, have been well developed and applied successfully in many application fields. However, for most classification learning algorithms that assume a relatively balanced distribution, the unbalanced class division of the data set encounters serious difficulties. 15,28,46,66 The characteristic of unbalanced data is that some classes have more occurrences than others. Since rare cases rarely occur, the classification rules for predicting small classes are often rare, undetected, or ignored. Therefore, test specimens belonging to small classes are more likely to be pre-classified than test specimens belonging to general classes.

In some applications, the correct classification of samples in a small class is usually more valuable than the opposite. For example, in the disease diagnosis problem compared to normal people, disease cases are usually very rare and the goal of identification is to detect people with the disease. Therefore, a favorable classification model is a model that provides a higher degree of recognition in disease categories. Therefore, the unbalanced or skewed class distribution problem is also called the small class or rare class learning problem.[1]

The study of imbalance problems is very important in data mining and machine learning. Two observations illustrate this point: (1) Class balance problems are common in a large number of important areas of the data mining community. Many research papers dedicated to this problem can be found in the references. 17 and other publications. The research work focuses primarily on three aspects of the class imbalance problem: (1) the nature of the class imbalance problem (i.e., in which areas does the class imbalance have the greatest influence on the performance of the standard classifier? "40); (2) solution to the class balance problem Possible solutions to the imbalance problem; (3) Appropriate measures for evaluating classification performance in the presence of class balance problems. This article describes the classification of unbalanced data. Every The following section studies an aspect of the imbalance problem - including the nature of the problem, the learning difficulty of the standard classification learning algorithm, the field of application, the learning objectives and the evaluation measures, the reported research solutions, the existence of several classes of imbalance problems. Finally, there are conclusions and some suggestions for future research.

**Application Domains**-The problem of class balance is common in a large number of fields that are very important to the data mining community. This problem lies in certain application areas. In some cases, this happens when the data collection process is restricted for some reason. 15 The following example illustrates this situation.

**Fraud Detection**- Fraud, such as credit card fraud and cell phone fraud, is a costly issue for many business organizations. In the United States, mobile fraud causes hundreds of millions of dollars in losses to the telecommunications industry each year. 70.80 The company tries to detect fraud by analyzing various consumption patterns in its transaction database. However, in their transaction collection, there are many more legitimate users than examples of fraud.

**Medical Diagnosis**- The clinical database stores a large amount of information about patients and their medical conditions. The data mining technology applied to these

© 2021 IJSRET

databases attempts to detect the relationship and patterns between clinical and pathological data in order to understand the progress and characteristics of certain diseases. The discovered knowledge can be used for early diagnosis. In clinical databases, cases of disease are quite rare compared to the normal population. As network-based computer systems play an increasingly important role in modern society, attacks on computer systems and computer networks have become more and more common. Learning prediction rules from network data is an effective method of detecting anomaly that can automate and simplify the manual development of intrusion signatures. It was found that there were different types of cyber attacks in the collection of network connection records - some were overwhelming, some were rare. For example, KDD-CUP'99 competition data includes four types of network attacks: denial of service (dos), monitoring (probe), remote to local (R2L), and user to root (u2r). Among these four types of attacks, u2r Categories R2 and R2 are inherently rare.[2]

Find oil spills from radar images of the sea surface. 46 Kun ca. 10% of oil spills come from natural sources, such as the seabed. More common is pollution caused by ships deliberately disposing of residual oil in the tank. Radar images from satellites provide the ability to monitor coastal waters. As oil slicks are reflected lower on the radar than ordinary sea surfaces, they look darker in the image. An oil spill detection system based on satellite imagery can be an effective early warning system that can deter illegal dumping of waste and can have a significant impact on the environment. Although satellites continue to produce images, there are far fewer images that contain oil spills than images without oil spills.

**Modern Manufacturing Plants.**For example, in a modern production plantOn the Boeing assembly line, 66 more and more processes are handled by automated or semi-automatic devices. Each device has a computer as its controller, which emits an automatic alarm when a defective pattern is detected. When building an alarm system through supervised learning, the number of available failures is significantly less than the number of regular programs.

In addition to these examples, other reported applications involve text classification 10 and direct marketing. Some of these applications, such as detection of fraud, detection of intrusion, medical diagnosis, etc. are also considered as problems with detection of anomaly. In anomaly detection, the goal is to find objects that differ from most other objects because anomalous objects and normal objects can be seen as defining two different categories, and a significant subset of anomaly detection systems treat anomaly detection as a dichotomy. Since anomalies are usually rare compared to normal observations, class imbalance problems are inherent with anomaly detection applications.

## II. NATURE OF THE PROBLEM

In the dataset with the problem of class balance, the most obvious function is the skew in the data distribution between classes. Skewed data distribution is not the only parameter that affects the modifier of the classifier that recognizes rare events. Other influential facts include the small sample size, separability, and existence of sub-concepts within the class.

## III. LITERATURE SURVEY

**Ali A. Alani et.al (2020)**in a smart home, the data generated by the sensor in real time is complex, noisy and incapable of balancing the data generated by the real-time sensor for human activity. It is a great challenge to create a machine learning model that can categorize activities that are not as common as other activities. Machine learning models designed to categorize unbalanced data typically learn more general categories. This learning bias occurs naturally, because the model is able to learn better classes with more records. This article investigates whether multimodal global data analysis can improve the classification results of nonlinear data compared to the use of static data; and compares the handling of unreasonable multimodal sensors when using different modification methods and deep learning models.

The experiments were performed using a large number of large-scale statistical data generated from the field of sensors for health (SPHERE) in the living environment. The data consists of 16104 samples, each with 5608 features, of one of 20 activities (classes) 20. The experimental results using SPHERE proved the challenge of coping with the data. a large number of unbalanced and highlighted the importance of having an appropriate variety of samples per class to train and test the in-depth learning model. In addition, the results show that when combining data and using SMOTE technology to correct for category imbalances, CNN-LSTM has the highest rate of 93.67%, followed by of CNN, 93.55% and LSTM, which was 92.98%.[1]

**Behzad Mirzaei et.al (2020)**in the problem of classification, the accuracy of classification will have a significant impact on training data. However, the distribution of data established in practical applications is not balanced. Unbalanced data means that most of the samples are in one class called the majority class, and there are a few samples in another class called the minority class. In these cases, most classifiers face this problem because they aim to classify samples that are evenly distributed between classes. Therefore, the selection of an appropriate training series is an essential step in the field of unbiased data classification. In this article, a proposal of an informative and effective technique is proposed to use the well-known DBSCAN algorithm to select the most suitable majority category.

© 2021 IJSRET

According to this algorithm, the most suitable samples were selected from the majority classes, and the samples from the other majority classes were removed to balance the training set. Compared to the other six preparation methods, the test results of more than fifteen unbalanced data sets prove the superiority of the proposed method.[2]

**Wei Wang; et.al (2020)** Traditional classification algorithms are based on uniform distribution of samples, and the results are inaccurate when processing this type of data. The main achievement is that the classification results are in the majority category. . Therefore, we present a non-standard multi-source heterogeneous data classification algorithm in this article, which is mainly based on the extension and extension of the supporting machines. Considering complex relationships in multiple data sources, presenting them as a logical, concise and efficient mathematical model can fully retain data information and improve the efficiency of data processing. We represent tennis and remove mining with heterogeneous data. This paper presents two different classification algorithms. In the first method, we directly represent the multiple heterogeneous data into a coherent torque form, obtain high-quality data through a dimensionality reduction algorithm, and then apply data classification by of the strengthening of the machine. Alternatively, we extract data from a variety of data sources and categorize them by a deep support vector machine (DSVM), which combines three DSVMs with the function of kernel. These algorithms are compared to a CUAVE data set containing two different audio and image types.[3]

J**aewoong Kang et.al (2020)** the problem of data imbalance is a major obstacle in the field of radiology. Although the accuracy of the categories is high due to the amount of data in the majority class, the compiler does not fully reflect the identity of the minority class. In this article, we propose a low-level modification method, which can accurately classify minority classes by changing the standard to approximate minority classes to majority classes. . This method makes repeated samples of a majority class with the same size as the minority class, but moves the standard by minimizing the influence of the data on the sample. near the cutting limit. In this way, the problem of imbalance in the field of radiology is solved in part, making it possible to study the discriminators using all the expensive medical imaging data.[4]

**Jun-Hai Zhai Etc.** . Study the problem of unbalanced data classification. Under the category of binary unbalanced data classification, this paper proposes a three-step classification method for large unbalanced data. Specifically, the first step uses the K-algorithm to compile large amounts of data into K clusters on a Hadoop platform. In the second step, we use the sample selection method to select significant samples from each cluster at the same time, and obtain the negative K rating. In the third step, we construct a balanced K training class in which the classes are located. -bad class and good class class, then we train K classifiers, and finally we combine these categories to classify the missing samples. Some experiments were performed to compare the proposed method and the two-of-art method with G-means. The experimental results show that the proposed method is more efficient and effective than the comparison method.[5]

**Baofeng Yao et.al (2021)** proposed a sub-sample classification algorithm for unbiased data (UECMS) based on clustering of mean changes for low-efficiency problems in small classifications numerical due to data inequality. The UECMS method uses clustering and sample selection to fill the sample under the majority of the sample. Selected samples and minority samples from the original data set create a new balanced data set. In addition, a bag-based mixed learning algorithm is used to categorize the balanced data. The experimental results show that the UECMS method improves the classification of minority classes in the non-statistical data.[6]

**Salma El Hajjami et.al (2020)**with the application of learning algorithms to practical applications, the problem of class imbalance has become more prominent. It has gained a lot of attention in the machine learning and data learning community. This problem exists in fraud screening, healthcare and many other fields, where training data is more representative of one category (called a majority category) than another category (called a minority category). number). Machine learning techniques try to deal with unbalanced data by focusing on mitigating the error rate in the majority class and underestimating the minority class. This is the most interesting from a learning point of view and also requires an expensive expense when the classification is not good. However, when learning from unbalanced data, the rate of imbalance is not the only reason for poor performance. Another major factor that accompanies the unbalanced data in the real world is the multiple occurrences of the two classes overlapping in the dominance space.

This problem is often referred to as class overlap, and we call it “behavioral noise”. In this article, we propose a way to reduce behavioral noise (OSBNR) to address the problem of class imbalance when there is a high level of behavioral noise. OSBNR is based on two phases. First, clustering is applied in the same case from a small group of individuals to a multi-behavioral cluster. Second, we select and eliminate cases of the majority category, which are perceived as behavioral noise and conflict with minority class achievement groups. [7]

**Xun Dong et.al (2020)** Intelligent Diagnostics is an important method of diagnostic identification of mechanical faults in the era of large-scale industrial data. Deep networks have received a lot of attention in this field due to their emergence of automated learning and the classification of input samples. The mixed-fuel network, in a conventional depth study model, was applied to the machine’s wrong steering. However, the limitation is that the input sample must be balanced to achieve an

© 2021 IJSRET

acceptable rate. In the mechanical operation process, there are more common samples and fewer samples with defects. Therefore, when using a bill-related network to process unbalanced data, the acceptance rate of a few categories is very low. Aiming at the above problems, an intelligent method of classifying unbalanced data based on a network corresponding to a depth value is proposed. According to this model, first of all, it studies the characteristic properties of the original mechanical signals through a number of modification and reinforcement operations. Second, it attaches these properties to mechanical health through fully connected layers. And finally, the purchase price loss gives an incorrect address to the different classifications of all the categories and updates them regularly during the training process to effectively classify the unbalanced mechanical data. The proposed method is validated by statistical analysis and processing data with different imbalance rates and compared with other methods. The experimental results show that the method is robust and can accurately categorize the unbalanced mechanical data.[8]

The data sharing of **Zonghai Zhu et.al (2020)** poses difficulties in classifying minority and majority samples of inequality problems. The balanced basket makes multiple samples below the maximum sample size, and combines the selected majority and minority samples to create a balanced average, which is the minority and minority samples. the majority is equal to the number. However, the balanced suitcase does not have a combined learning platform. In addition, it ignores the correlation of all subsets and global information about all data distributions. For this reason, this article places a multi -pronged approach to becoming an effective learning system with regular practice. In the study area, the so-called RS control code establishes a relationship and realizes the co-operation of all dependencies by requesting a constant output from a few samples of the fragment -different things. In addition, another code called RW provides global information for each base category by minimizing the difference between the solution vector direction per subset and the path of the solution vector in the set data. The proposed learning plan is called Globalized Multiple Balanced Subset with Collaborative Learning (GMBSCL). The test results confirm the effectiveness of the GMBSCL proposal.[9]

**Qibin Wang et.al (2020)**Due to the unbalanced distribution of highly infrared spectroscopy data, it is difficult for traditional machine learning methods to accurately classify the samples in the modeling process. In general, near-infrared data spectroscopy data are high quality and the samples are few.

To improve the accuracy of machine learning classification, we present an ensemble -based learning method here. Specifically, the proposed method uses a primary Variational Autoencoder (VAE) network to generate multiple samples, and combines these samples with the original data to create a new data set. Then use the multi -type cascading forest fusion method (FCForest) to construct a classification model. We validated and evaluated our method by means of a noninvasive near -infrared smoke spectroscopy from citrus greening. The experimental results show that the use of VAE to create samples improves the accuracy of the classification of unbalanced data. In addition, by using the FCForest method to smoke balanced new data, the performance of the categories is improved.[10]

**Imbalanced class distribution**

For the two-class problem, the imbalance in the class distribution can be expressed by the ratio between the plug size of the small class and the size of the popular class. In practical applications, the ratio can reach 1: 100, 1: 1000 or even greater. 15 For reference. 85. Research was conducted to examine the relationship between the class distribution of the training dataset and the classification performance of the decision tree. Their research shows that a relatively balanced distribution usually gives better results. However, it is not clear how unbalanced the class distribution will reduce the classification performance because other factors such as sample size and separability will also affect performance. In some applications, a ratio as low as 1:35 may make some methods insufficient to build a good model, while in other cases 1:10 is difficult to handle. 41

**Small sample size-** The problem of class balance, when the degree of imbalance is fixed, the sample size plays an important role in determining the "good or bad" of the classification model. In the case of limited sample size, it is unreliable to disclose the inherent laws in small classes. Experimental observations reported in references. 40 shows that as the size of the training set increases, the large error rate caused by the unbalanced class distribution decreases. This observation is completely understandable. When more data is available, relatively more information about small-class classification modeling is facilitated, so that rare samples can be distinguished from most samples. Therefore, the author of the reference. 40 shows that if a large enough data set is provided, the unbalanced class distribution may not be an obstacle to the classification. It is assumed that the dataset is available and the required learning time for a fairly large dataset is acceptable.

**Class separability**-The difficulty of separating subcategory from popular category is the key issue with subcategory issues. If there is a very discriminatory pattern between each class, there is no need for very complex rules to classify objects. However, if the patterns between each class overlap at different levels in particular functional spaces, it is difficult to impose discriminatory rules. Experiments performed in references. 60 Changing the degree of overlap between classes leads to the conclusion that “the very imbalance of the class itself does not seem to be a problem, but when associated with very overlapping classes, it can significantly reduce the number

© 2021 IJSRET

of correctly classified few (small ) class examples. "A similar statement based on the experiment reported in the reference. 40 is" The linearly separable domain is not sensitive to any number of imbalances. In fact, the sensitivity of the system to imbalances increases as the complexity of the concept increases. "

## IV. METHODOLOGY

In this section we focussing on various data level approaches that help to balance the data and eliminate the class imbalance problems. We were also discussing our proposed methodology to solve the class imbalance problem. There are sections (A) that explain various approaches to handle class imbalance problem (B) Proposed methodology.

**1. Approaches to handle class imbalance problems:**
Data – level approach as resampling technique is used or a pre-processing process to rebalance the class distribution. Resampling is a common technique for this approach. The main goal is to gain more data from minority class. So, there is two approaches namely Undersampling and Oversampling used in it. If we removed the instances from majority class to balanced the datasets its called undersampling and if we add the similar instances in minority class to balanced the ratio of the class then its called oversampling. Now we see different approaches below in detail.[1]

**2. Undersampling:** Itsan technique to balance the class distribution for classification of datasets.In imbalance classes are categories in 2 classes one is majority class and another is minority class. Under sampling basically reduced the data simply from consists datasets by eliminating some examples from majority class to balance the objectify to equalizing numbers to other given class. This method used on training datasets. Basically under sampling methods coordinate with oversampling technique of minority class and both this techniques combines and give better performances on training datasets. the random selecting is simple under sampling technique which selects the examples from majority class and deleting from training datasets.

**3. Near Miss-** Its an under-sampling technique. Instead of resampling the Minority class, using a distance, this will make the majority class equal to minority class. There are three near miss methods

**4. Random under sampling-** Removes samples randomly with or without replacement from the majority class. This is one of the preliminary methods used to eliminate the imbalance in the dataset, however, it may eliminate useful or important samples as it increases in the variance of the classifier.

**5. Oversampling-** This method used to reduction in minority class so that the there is an equal and balanced distribution in class. SMOTE is an commonly recommended in oversampling.

**6. ADASYN-** adaptive synthetic sampling approach expands on the procedure of SMOTE by shifting the importance of the classification boundary to those minority class which are difficult. ADASYN uses a weighted spread for not similar minority class examples as per their level of complexity in learning, where more synthetic data is created for minority class examples that are difficult to learn.

**7. SMOTE:** There are a number of methods available to oversample a dataset used in a typical classification problem (using a classification algorithm to classify a set of images, given a labelled training set of images). The most common technique is known as SMOTE: Synthetic Minority Over- sampling Technique. To illustrate how this technique works consider some training data which has s samples, and f features in the feature space of the data. Note that these features, for simplicity, are continuous.

As an example, consider a dataset of birds for classification. The feature space for the minority class for which we want to oversample could be beak length, wingspan, and weight (all continuous). To then oversample, take a sample from the dataset and consider its k nearest neighbours (in feature space). To create a synthetic data point, take the vector between one of those k neighbours, and the current data point. Multiply this vector by a random number x which lies between 0, and 1. Add this to the current data point to create the new, synthetic data point. Many modifications and extensions have been made to the SMOTE method ever since its proposal.[7]

## V.PROPOSED METHODOLOGY

In given figure 1 there is proposed system in which various approaches are used to balanced the class. Sampling techniques are used to performed by adding and removing the instances of classes. Classifiers mainly perform to balance out the data. Training and Testing approaches are done by which we found how much of data is rectifying by which we finally applying our algorithms to find better performance and average of all techniques which is going to be used in it.

© 2021 IJSRET

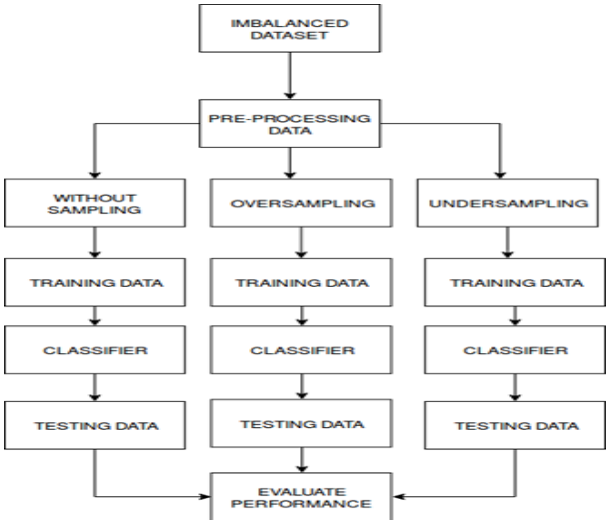


Fig. 1 Proposed Methodology.

## VI. CONCLUSION

This review conducts a comprehensive study of recent research advances in the classification of unbalanced data. This article begins with the examples of several application fields that are disturbed by the problem of imbalance in the class, and discusses the essence of the class imbalance problem; reviews most of the standard classification learning algorithms, such as decision trees, backpropagation neural networks, Bayesian networks, recent neighbors, supports vector machines, and association classification to gain insight into their learning difficulties when encountering unbalanced data; a comprehensive study of the reported research solutions to this problem has been conducted and their benefits and limitations have been studied with a view to Inspire advanced research ideas in the future.

## REFERENCE

1. Ali A. Alani; Georgina Cosma;Aboozar Taherkhani Classifying Imbalanced Multi-Modal Sensor Data For Human Activity Recognition In A Smart Home Using Deep Learning 2020 International Joint Conference On Neural Networks (IJCNN) Year: 2020 DOI: 10.1109/IEEE Glasgow, UK
2. Behzad Mirzaei; Bahareh Nikpour;Hossein Nezama badi-Pour An Under-Sampling Technique For Imbalanced Data Classification Based On DBSCAN Algorithm 2020 8th Iranian Joint Congress On Fuzzy And Intelligent Systems (CFIS) Year: 2020 DOI: 10.1109/IEEE Mashhad, Iran
3. Wei Wang; Mengjun Zhang;Li Zhang;Qiong Baiim Balanced Data Classification For Multi-Source Heterogenous Sensor Network IEEE Access Year: 2020
4. Jaewoong Kang ; Mye Sohn Recursive Undersampling-Based Decision Boundary Alignment For Imbalanced Radiology Image 2020 IEEE International Conference On Big Data And Smart Computing (Big Comp) Year: 2020
5. JUN-HAI ZHAI;SU-FANG ZHANG;MO-HAN WANG;YAN LI A Three-Stage Method For Classification Of Binary Imbalanced Big Data 2020 International Conference On Machine Learning And Cybernetics (ICMLC) Year: 2020
6. Baofengyao;Lei Wang An Improved Under-Sampling Imbalanced Classification Algorithm 2021 13th International Conference On Measuring Technology And Mechatronics Automation (ICMTMA) Year: 2021
7. Salma El Hajjami;Jamal Malki;Alain Bouju; Mohammed Berrada A Machine Learning Based Approach To Reduce Behavioural Noise Problem In An Imbalanced Data: Application To A Fraud Detection 2020 International Conference On Intelligent Data Science Technologies And Applications (IDSTA) Year: 2020
8. Xun Dong;Hongli Gao;Liang Guo;KesiLi;Andongzhe Duan Deep Cost Adaptive Convolutional Network: A Classification Method for Imbalanced Mechanical Data IEEE Access Year: 2020
9. Zonghai Zhu;ZheWang; DongdongLi;Wenli Du Globalized Multiple Balanced Subsets With Collaborative Learning for Imbalanced Data IEEE Transactions on Cybernetics Year: 2020
10. Qibin Wang; LingqiaoLi; Xipeng Pan; Huihua Yang Classification of Imbalanced Near-infrared Spectroscopy Data 2020 12th International Conference on Advanced Computational Intelligence (ICACI) Year: 2020

© 2021 IJSRET